DOCUMENT RESUME

ED 222 516 TM 820 661

AUTHOR Smith, Philip L.; Teeter, Phyllis Anne
TITLE The Use of Generalizability Theory with Behavioral Observation. Draft.
PUB DATE Mar 82
NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavior Patterns; *Behavior Rating Scales; Clinical Psychology; Multivariate Analysis; *Observation; Psychological Evaluation; Reliability; *Research Methodology; *Statistical Analysis; Statistics; Validity
IDENTIFIERS Behavioral Measures; Covariation; *Generalizability Theory

ABSTRACT

Behavioral observation has become increasingly important in research in developmental and clinical psychology, as well as in program evaluation and instructional research. Several statistical approaches to assessing the dependability of behavioral observation were reviewed. The demand for accurate and efficient paradigms for behavioral observation has increased as psychologists view the method as a more viable alternative to traditional psychometric devices. Generalizability theory removes some ambiguity in interpretation by explicitly defining a universe of admissible observations and a universe of generalization. Analogies to traditional statistics were made using generalizability theory statistics to illustrate the similarity of concept and less ambiguous interpretation. Multivariate approaches were suggested for use with composite scores and situations where covariation between levels of a facet are of interest. Various approaches that have been suggested for behavioral observation were summarized. Continuous type scores were the primary focus of this study. More research is needed on agreement statistics when the nature of the task involved is classification on a nominal scale. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222516

DRAFT (3/9/82)

The Use of Generalizability Theory with Behavioral Observation

Philip L. Smith
Phyllis Anne Teeter
University of Wisconsin-Milwaukee

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. L. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820-661

Paper presented at the Annual Meeting of the American Educational Research Association, New York. March 22, 1982.

Abstract

Behavioral observation has been playing an increasingly important role in research in developmental and clinical psychology as well as in areas such as program evaluation and instructional research. Problems in the construction and use of behavioral observation paradigms have been well documented, and it is largely due to these problems that behavioral observations have not been used more extensively in the past. As psychologists begin to view behavioral observation as a more viable alternative to traditional psychometric devices, the demand for more accurate and efficient paradigms for behavioral observation has emerged. The purposes of this paper are to (1) summarize various approaches that have been suggested for behavioral observation, (2) review issues in the assessment of the technical adequacy of behavioral observation paradigms, (3) describe the important role that generalizability theory can play in the construction and use of a behavioral observation system.

In addition to its traditional and major role in developmental and clinical psychology, observation as a behavioral assessment technique is gaining an increasingly important role in program evaluation and instructional research. Frequently, behavioral observations are thought to be more complex and subjective, and therefore less accurate, than other traditional assessment procedures. While this is true for many observational systems primarily due to their lack of standardization, this need not be the case. It is true that observational approaches are more complex since they require the observer to observe an event and then record this event in some accurate, perhaps quantitative form. Adding to the complexity of the task is the fact that the behavior of interest is often less well defined with fewer operational parameters than is desirable for accurate, reliable assessment. Critical variables in the observational procedure are the observer, the conditions under which the observation is made, the object of the observation, the recording procedure, and the translation and interpretation of the observation. While analogies can be drawn between these 'high inference' observational procedures and more traditional 'low inference' (e.g., paper and pencil type tests and scales) assessment procedures, standardization throughout the latter type procedure plus their inherent objectivity serve to limit or reduce the variability of

interpretation in each of the critical variables mentioned above.

Because of the high inference nature of behavioral observations, their accuracy and consequently their usefulness has been questioned on several grounds (see for example Finn, 1972; Taplin & Reid, 1973; Ghiselli, 1963; Mash & McElwee, 1974; and others). The more persistent of the supporters see the value of observation, despite its technical complexity, as a more accurate and ecologically valid way of recording the behavior of an individual or group. Clinicians have traditionally relied heavily on observational techniques, while other psychologists and educators are quickly adopting more observationally oriented techniques hoping to provide a more wholistic or multidimensional picture of the behavior of interest. While this trend does not obviate the concern researchers express for the complexity of the observational task, it does point to an increased need to consider the technical adequacy of such procedures.

The purpose of this paper is twofold. First, a brief review of approaches addressing the question of the technical adequacy of behavioral observations will be presented. Second, the technical concerns identified and discussed will be examined from the perspective of generalizability theory. The variety of techniques discussed are applied to a common data set for comparative purposes.

Technical adequacy: Traditional perspective

The following is a brief review of approaches that have been suggested for assessing the technical adequacy of behavioral observations. For a more exhaustive review, the interested reader is referred to Tinsley and Weiss (1975) or Mitchell (1979).

Validity. Many of the theoretical foundations and classical concerns regarding the validity of measurements, including observations, are summarized by Cronbach (1963). With observational data, questions of validity are complex, but researchers are normally concerned with only two aspects: (1) explicitly defining the behavior to be observed, and (2) accurately translating this observation to the required form. The first of these aspects is generally addressed by operationally describing the behavior of interest to the observer. A complete and unambiguous description will increase the likelihood of the observer attending to and recording the intended behavior. This behavioral description may be equated with the content validity of observation.

The second aspect of validity is concerned with the method in which the target behaviors are recorded. Methods of recording observations include checklists, frequency counts, and anecdotal records, the choice of which depends upon how the specific behavior of interest is theoretically defined. The way in which observations are recorded has implications for both the content and construct validity of observations. The validity of

observational procedures, in the classical sense, has been more concerned with form than any type of technical assessment. While more quantitative techniques such as the multitrait-multimethod techniques introduced by Campbell and Fiske (1959) could be applied to validate the observational process, normally they are not utilized because alternative indicators of the critical behavior are often difficult and expensive to obtain.

Reliability. The concept and nature of the reliability of behavioral observations has a reasonably extensive history, however, it has received only nominal attention from clinicians and practitioners. Discussions of reliability focus primarily on methodological issues such as the type of scores one derives from an observational procedure, and on the type of reliability one is interested in obtaining.

The type of score derived from an assessment procedure can have a profound effect on how that score is treated in assessing reliability. For example, some observational schemas involve simple nominal classification (without reference to order) resulting in a qualitative assignment of the measured behavior. In a circumstance where a psychologist is interested in hyperactivity, the observer might observe a subject and classify specified behaviors as either hyperactive or not hyperactive. Other types of observational procedures might require the observer to rate behavior on some specified construct (e.g.,anxiety or hyperativity) on a continuum (e.g.,high/low anxiety or hyperactivity), resulting in an ordinal scale measurement. Interval and ratio

scales are also used and are most evident when the observational schema calls for a frequency count of some behavior. For instance rather than simply rating an individual as being either high or low on a scale of anxiety, an observer might record the number of times a subject demonstrates anxious behavior in a specified time frame or observation session. Wiggins (1973) characterizes these types of observational procedures as category systems, numerical rating scales, and sign systems respectively. While Wiggins also discusses other observational and rating systems, the three mentioned above encompass those most commonly used. Frequently the operational definition of some behavior or trait in question is related to the type of scale used. As the recording moves from a nominal (category) system to a ratio (sign) system, the task of observation becomes less inferential and consequently the behavior becomes more concretely and explicitly defined. Classical measurement theory recognizes that reliability is affected by the way behaviors are classified and recorded. Also, classical measurement paradigms create the need for several types of reliability. Classical measurement theory differentiates between intrarater and interrater indicators, order and agreement indicators, and consistency and stability indicators of reliability. Again the type of score derived determines how these specific reliability indicators are methodologically defined and interpreted.

Intrarater vs. Interater Reliability. This basic distinction calls attention to the difference between the correspondence

of two or more ratings made by the same rater (intra), versus a correspondence of ratings made by two or more observers (inter). The way in which a particular observational scale is utilized and the kinds of behavior being studied help determine whether interater or intrarater reliability is most relevant in a given circumstance. An index of intrarater reliability is initially important in the development of any observational procedure or scale because of the need to reduce the amount of variability within individual raters when the procedure is applied several times on the same behavior sample. That is, the observational procedure (or an item within a scale) must be defined in such a way that it provides the rater with a cognitive or operational framework that insures consistent ratings across several observations of the same behavior. For example, if a psychologist is interested in assessing hyperactive or disruptive behaviors in the classroom it is necessary for the psychologist to apply the same criteria for judging which behavior is to be classified as disruptive, and, it is also imperative for the criteria to be applied consistently across several ratings of the same behavior to insure high intrarater reliability.

While intrarater reliability measures consistency <u>within</u> individual raters, interater reliability provides a measure of consistency <u>between</u> raters. Using the previously cited example, an index of interater reliability reflects the degree of agreement between two raters (e.g., psychologists) as to what constitutes a given level of hyperactive or disruptive behavior. So, a

high interater reliability would indicate that the scale or procedure produces relatively uniform ratings, and that there is some concensus across observers as to the behavior which occurred. A measure of interater reliability is called for when this correspondence is necessary and is particularly relevant when several observers are used in some manner to rate members of a group or when ratings must conform to some type of standard. Although it is seldom noted, it is obvious from the above distinction that intrarater reliability is a requisite for interrater reliability. This is an analogous situation to the reliability coefficient serving as a theoretical upper limit to a validity coefficient. Interrater reliability, in theory, can only be as high as intrarater reliability.

Order vs. Agreement Measures. Reliability assessment implies an attempt to reproduce a set of scores through maximally similar procedures. Normally, two sets of scores are correlated to provide evidence of this reproducability (Ebel, 1951). The correlation coefficient derived from such a procedure is of value only to the extent that the set of observations maintain the same relative order upon reproduction. If two observers rank individuals (or sessions or events) relatively the same, regardless of the absolute values of the scores, the data are thought of as being reliable. For example if two psychologists observe five children in a classroom and assign each a score (from 1 to 10) on a scale measuring disruptive behavior, a high correlation coefficient (order) would be derived if the judges rank the children in

the same relative order. In this instance a comparative statement can be made concerning the relative amount of disruptive behavior, so that, we can discern from these data that more than one observer ranks the behavior of child A as more disruptive than the behavior of child B. When comparative decisions of this nature are of major concern, a reliability coefficient measuring order is satisfactory. However, if the absolute score value is clinically important (e.g., has the hyperactive behavior of child A been ranked high enough), or if we are interested in determining if one psychologist can reproduce the exact score (not only its rank order) then a measure of agreement is needed (Light, 1971). Again using the previously described example, even though child A may have been ranked as the most disruptive in the sample by both psychologists, one psychologist may have assigned child A a value of 10 on the scale measuring severity of disruption while the other may have assigned a value of 6. The high correlation obtained when rank order was of primary concern may be insufficient in circumstances where rater agreement is of concern. For this reason, indices of observer or rater agreement have also been proposed, and the form of which depends upon the type of measurement scale used. Some of these indices are briefly discussed in a later portion of this paper.

Stability vs. Consistency Indices. Distinctions between stability and consistency measures of reliability are typically procedural, not statistical, and are more closely related to the nature of the specific behavior being assessed rather than the

research or clinical questions of concern. Measures of stability are designed to detect rating or behavior differences across time, and are therefore most relevant in longitudinal type studies. Consistency measures are concerned with the accuracy of a rating at a given point in time. The two measures need not provide redundant information. For example, an observational scale may consistently measure a child's level of hyperactivity during a single observational session but still lack stability. Because hyperactivity may not be an enduring behavioral characteristic, the stability of the trait is variable and may change depending on a particular intervention, a measure of stability would be inappropriate for evaluating the reliability of the particular observational scale. Correlational approaches are universally applied to determine both stability and consistency with the major distinction between the two resulting from differences in the amount of time provided between observational sessions.

Because of the flexibility in methods for gathering and recording observational data, a number of different indicators of reliability for rating and observational scales have been generated. Tinsley and Weiss (1975) provide a fairly comprehensive review of these procedures, several of which are outlined below.

In general, traditional indices of reliability for observational scales are classified on the basis of whether agreement between raters is most important or whether raters are consistent when rank ordering subjects on some behavioral dimension. With

nominal (classification) measures, rank ordering is not meaningful consequently there are no statistics for this circumstance. For ordinal data, Finn (1970) has suggested using an interrater index that reflects the proportion of chance variation in a set of ratings by comparing actual within subject variation with that expected by chance. For interval data, an intraclass correlation is often suggested as an appropriate measure of rater consistency (Ebel, 1951, Guilford, 1954, Bartko, 1966). Traditional intraclass correlation compares total variance to between subject variance in a rater x ratee ANOVA, thus reflecting the proportion of total variance due to true ratee (subject) differences. This intraclass correlation may be adjusted for the number of ratings using concepts similar to those provided by the Spearman-Brown adjustment.

Due to the inappropriateness of order statistics for nominal data, and because rater agreement is attracting more attention, several statistical methods have been developed for indicating rater agreement for nominal data. These developments range from proportion of agreement statistics, including chi-squares and Kendalls rank order statistics, to the more contemporary Kappa statistics (Cohen, 1960, 1968; Fleiss, 1971; among others) and varieties.

Contributions from Generalizability Theory

Regardless of the statistical index used to describe their technical aspects, behavioral observation schemes and rating scales have been the subject of considerable criticism, espe-

cially regarding the adequacy of their reliability assessment (see, for example, Ghiselli, 1963 or Finn, 1972). Much of the criticism is due to the failure of investigators to adequately specify the full conditions in the measurement situation that affect reliability, including the procedure's insensitivity to its inherent multifactored nature. The fact that various reliability indices for different observational/rating circumstances have been suggested serves to further substantiate this point. The judgement and/or interpretation regarding the adequacy of an instrument for the observation or rating of specific behaviors, and the degree to which the procedure accomplishes its stated purposes is closely tied to the reliability index that is selected.

It has been suggested here and elsewhere (Cone, 1977) that Generalizability theory provides a viable alternative to the classical-traditional procedures for determining the technical adequacy of observational measures, particularly for ordinal and interval scaled data. For categorization schemes, the properties and behavior of the generalizability theory approach has not been well studied, and therefore is not recommended. When a system involves two levels of classification (i.e., dicotomous classification), the generalizability coefficient computed on dicotomous scores might be an adequate substitute. For multicotomous nominal scales, the procedure is clearly inappropriate although analogies using log-linear models are currently being studied. With respect to ordinal or higher order observational data, there

are at least two principle advantages to using the generalizability theory framework: 1) the generalizability study forces the investigator to specify a universe of scores to which he would like to generalize; and, 2) appropriately collected generalizability study data can be construed in a number of ways (for decision studies) to provide data regarding questions of reliability that would otherwise require multiple indices in the more traditional paradigm.

The use of generalizability theory has served to clarify much of the confusion surrounding the term "reliability coefficient" used in conjunction with measurement in general. Many of the advantages of applying generalizability theory to classroom observations are discussed in detail by McGaw, Wardrop, and Bunda (1972) while Wiggins (1972) provides a fairly comprehensive discussion in reference to behavioral observations. A principle advantage of the approach in both situations is that the numerous sources of circumstantial variability (e.g., situations, occasions, conditions, observers) which formerly required that several sets of data and coefficients be constructed and reported are simultaneously considered, and the universe or population from which these are drawn is explicitly and unambiguously defined.

Essentially, generalizability theory uses an analysis of variance framework to partition the total variance of scores resulting from a measurement procedure into the various sources of variance (factors) which are explicitly considered in the

design. The variance attributed to the facet of differentiation (that factor in a design corresponding to the object of measurement) is compared to the observed score variance (that due to the facet of differentiation plus error) which forms an intraclass correlation coefficient similar to the traditional intraclass index discussed earlier. Because generalizability theory considers explicitly multiple sources of error (due to the factors considered), indicating where measurement errors exist and how they might be reduced in subsequent applications of the observational procedure, it is more appealing than the traditional intraclass index.

As a first step, generalizability theory forces the investigator to specify a universe of admissible observations for a measurement procedure. This universe includes an explicit description of those factors in the measurement procedure that may be allowed to vary from observation to observation. These factors (called facets) form the ANOVA design used in generalizability study and are considered in forming the error term for the generalizability coefficient. The variance attributed to each source is estimated using variance component estimates (Crump, 1946) from the expressions for the expected mean squares of the ANOVA design of interest. Once the initial data (called G study data) are collected, an intraclass correlation is formed by taking the ratio of universe score variance (that due to facet of differentiation) to observed score variance (that due to differentiation plus sources involving interactions with this term) to

form the generalizability coefficient. The notion of Spearman-Brown may be applied to this coefficient to project the adequacy of a procedure to a 'universe of generalization' which may be equivalent to or may be some subset of the previously defined universe of admissible observations.

For example, McGaw et al. (1972) in reference to classroom observation specify a design configuration that allows for several types of "reliability" coefficients. The G study design has four facets and crosses Teachers, Situations, and Judges and nests Occasions within situations. McGaw et al. discuss three possible coefficients that can be derived from this design: 1) that reflecting the extent to which teachers can be distinguished (teachers are the facet of differentiation); 2) that reflecting the extent to which situations can be distinguished; and, 3) that reflecting the extent to which teacher behavior changes from situation to situation (the interaction is the focus of differentiation). While slightly different from the McGaw et al. coefficients, these may be defined respectively as:

$$\hat{\rho}_t^2 = \hat{\sigma}_t^2/(\hat{\sigma}_t^2 + \hat{\sigma}_\delta^2) \qquad (1)$$

where $\hat{\sigma}_\delta^2 = \hat{\sigma}_{tj}^2/n_j' + \hat{\sigma}_{tsj}^2/n_s'n_j' + \hat{\sigma}_r^2/n_o'n_s'n_j'$

$$\hat{\rho}_s^2 = \hat{\sigma}_s^2/(\hat{\sigma}_s^2 + \hat{\sigma}_\delta^2) \qquad (2)$$

where $\hat{\sigma}_\delta^2 = \hat{\sigma}_{sj}^2/n_j' + \hat{\sigma}_{tsj}^2/n_t'n_j' + \hat{\sigma}_r^2/n_o'n_t'n_j'$

$$\hat{\rho}_{st}^2 = \hat{\sigma}_{st}^2/(\hat{\sigma}_{st}^2 + \hat{\sigma}_\delta^2) \qquad (3)$$

where $\hat{\sigma}_\delta^2 = \hat{\sigma}_{tsj}^2/n_j' + \hat{\sigma}_r^2/n_o'n_j'$

and where the $\hat{\theta}^2$ terms are variance component estimates derived from the above-mentioned ANOVA design and n'_j, n'_s, n'_t, and n'_o are the number of judges, situations, teachers, and occasions respectively to be used in the application of the procedure.

The first of these coefficients is the one most often associated with data of this nature and reflects the ratio of variance between teachers to that due to observed score differences. In this case, unlike the simple intraclass coefficient, error can be specifically attributed to various sources (see the equations for σ^2_δ) and subsequent prescriptions for changing the measurement procedures to reduce these errors can be made. For example, when teachers are the object of measurement if the teacher by judge interaction ($\hat{\sigma}^2_{tj}$) term is large, thereby reducing the value of the generalizability coefficient, one could sample more occasions (increase n'_j) in subsequent application of the procedure to obtain a more stable estimate ($\hat{\sigma}^2_{tj}/n'_j < \hat{\sigma}^2_{tj}$) of the behavior of interest.

McGaw et al. make a unique contribution in the presentation of their second and third reliability coefficients. In the second coefficient, situations are distinguished in the same manner that teachers were distinguished in the initial coefficient. Although this reliability would normally be assessed by a correlational procedure (if order is of concern) or some statistical test (if situation means are of concern) or both, the generalizability theory perspective can accommodate both of these

situations. For example, the generalizability coefficients above express relative ordering. In the first equation the variance components associated with occasions, judges, and situations and their interactions contribute to <u>systematic</u> error which does not effect relative rank order of teacher ratings and thus is not included in the error term, but which <u>will</u> affect departures from true score. Consequently, any statistic that is used to express this departure from true score <u>must</u> consider these sources. Generalizability theory defines an error $\sigma^2(\Delta)$ which includes in addition to error terms included in σ^2_δ those terms that affect departure from true score. For equations (1) and (2) these errors are

$$\sigma^2(\Delta)_t = \sigma^2_s/n'_s + \sigma^2_o/n'_o n'_s + \sigma^2_j/n'_j + \sigma^2_{ts}/n'_s + \sigma^2_{tj}/n'_j + \sigma^2_{sj}/n'_s n'_j + \sigma^2_{tjs}/n'_s n'_j + \sigma^2_r/n'_o n'_s n'_j \quad (4)$$

$$\sigma^2(\Delta)_s = \sigma^2_t/n'_t + \sigma^2_o/n'_o n'_t + \sigma^2_j/n'_j + \sigma^2_{ts}/n'_t + \sigma^2_{sj}/n'_j + \sigma^2_{tj}/n'_t n'_j + \sigma^2_{tjs}/n'_j n'_s + \sigma^2_r/n'_o n'_j n'_t \quad (5)$$

Although the use of this term in place of σ^2_δ in an intraclass correlation makes little interpretive sense (aside from expressing proportion of variance accounted for by persons), its value provides a convenient yardstick for assessing measure agreement. Golding (1972) addresses this issue in detail with respect to ω^2 statistics in ANOVA (which use $\sigma^2(\Delta)$). Golding's arguments are made in a context relating to circumstances where persons are the object of differentiation, but these arguments can apply equally well to circumstances where situations are the object of

differentiation. Golding's comments are aimed specifically at individual differences situationism research, where an ω^2 ratio has traditionally been used, and is therefore an index of agreement since variance due to situations has been considered error. Golding argues that if the intent is to generalize rank order, elimination of situation specific variance, and all variance not tied to the face of differentiation should be ignored, resulting in a generalizability coefficient (intraclass correlation) which reflects only rank order consistency. It is argued later that even when agreement is of interest, attention should focus on $\sigma^2(\Delta)$ not an intraclass coefficient based on $\sigma^2(\Delta)$, primarily for reasons of interpretation. Since $\sigma^2(\Delta)$ reflects absolute error, it provides an appropriate indication of how accurately a given score can be interpreted relative to its true score.

McGaw et al.'s third coefficient provides information on the differentiation of any teacher (person) by situation interaction. For our purposes this coefficient focuses on how well observers can detect behavioral differences in teachers from situation to situation, regardless of whether behavior is expected to change or not, consequently the relative error term (σ^2_δ) makes the most sense.

The McGaw et al. developments are important because many of the applications of G theory to observational data and behavioral observation pertain to classroom observations. Shavelson and Dempsey-Atwood (1976) have provided an excellent review of teacher observation studies that can be construed under a

generalizability theory paradigm. To paraphrase their premise, Shavelson and Dempsey-Atwood contend that past inability to link teacher behavior to student outcome may be a result of unstable measures of teacher behavior due to a) actual teacher changes, b) teacher changes due to situations, c) pupil differences and changes, and d) limitations in the measurement method. The authors recommend that generalizability theory be used to help solve these problems.

While the most extensive use of generalizability theory with behavioral observation has probably been done with teacher observation, few studies have been reported that use this framework for assessing the adequacy of these procedures. Because most studies of teacher behavior do a poor job of reporting measurement conditions, the Shavelson and Dempsey-Atwood study could not report firm conclusions regarding the stability or consistency of teacher behavior in the intended universe. These same conclusions would probably apply if one were to examine observational studies of behavior that has been less well investigated. Researchers are simply not used to adequately describing the observation situation sufficiently enough to formulate an appropriate universe, or to providing adequate data from which to form conclusions regarding the behavior of interest.

Few studies explicitly report using G theory for technical assessment, and its application to previous studies is limited in many instances because of inadequate description of the observational situation and because insufficient data were collected.

When appropriate data are collected and analyzed from a generalizability theory perspective, more often the data are reported for rating scales which use observation (see Levy, 1974 and Weiöer & Weiss, 1980). However, Wiggins (1972) and Cone (1977) have discussed aspects of behavioral assessment and observation which should be considered in determining the generalizability of such data. Wiggins' schema which is the more comprehensive of the two urges a consideration of the following facets in a G study of behavioral assessment: situation/settings, observers, instruments, occasions, and attributes.

The type of instrument used and the nature of the resulting score will normally be fixed (thus will not vary) for a given study. Here it is assumed that a single instrument is used for such observation and that the score derived is ordinal or higher. Often attributes will be fixed in the universe at one, or at least at a finite level in a given study. For the sake of simplicity and clarity, we will consider assessment strategies that focus on a single attribute only. The use of multiple attributes leads to several questions that can greatly complicate the G study design (see Smith and Kane, 1981).

Given these assumptions we are led to a design of the type suggested by McGaw et al. (1972), but is more generally applicable to behavioral assessments where the teacher facet is replaced with a more general 'subject' facet. This design, with observers are crossed with situations and subjects, and occasions nested within subjects will subsequently be referred to as design

1. There are other procedures which might require the observational situation to be fixed at one in the universe of admissible observations, consequently no generalization would be intended over situations to help standardize the procedure. In this circumstance a simpler design would be used where observers are crossed with occasions and nested within subjects. This design will be referred to as design 2. Design 2 should not be confused with a circumstance where situations are allowed to vary in the G study but then are fixed in a subsequent (D) study (universe of generalization). In this latter case (a variation of design 1), situations are assumed to vary in the universe of admissible observation and this source of variance must be explicitly considered in the D study.

The expected mean squares for the random effects case, and the computational procedures to obtain variance component estimates for each of these designs is shown in Table 1 and 2 respectively.

Insert Tables 1 and 2

About Here

A reasonable third design is one where judges (observers) are linked to occasions which are nested within subjects. That is, unique judge-occasion combinations are used for each subject. If judges are the same for all subjects, the linking procedure would suggest that the judge-occasion combinations

still remain nested in the design. Situations may or may not be present as a facet in this design. This design and its variations is not explicitly considered in this paper and is not recommended for G study application because too much information is lost regarding sources of variance. It may be, however, that this design configuration is the most appropriate for subsequent decision study application.

Returning to designs 1 and 2, occasions would ordinarily be considered random. In design 1, situations might be considered either fixed or random in the universe of generalization, resulting in four possible design circumstances as indicated in Figure 1. Design 2, where the situation is fixed in the universe of admissible observations leaves us with two common design possibilities, one where raters are fixed and the other where raters are random. The various coefficients that can be derived from these designs and their relationship to more traditional indices are indicated below.

Insert Figure 1

About Here

Design 2. Because of its simpler nature design 2 is considered first. This design includes in its universe of admissible observations facets associated with observers (B), persons (A), and occasions (D). Situations are standardized across all applications. From this basic design we can distinguish two varia-

tions, Design 2A and 2B, which differ in terms of their universe of generalization. In design 2A, observers are considered random while in design 2B no generalization is intended beyond those included in the D study and observers are thus considered fixed. In most local applications, this latter design variation is often the more reasonable, but this distinction is seldom considered and not accounted for in traditional reliability statistics. When persons are the object of differentiation, the following coefficients are appropriate where the observer is considered 'random' and 'fixed' respectively:

$$\hat{\rho}^2(B,D) = \frac{\hat{\sigma}^2_a}{\hat{\sigma}^2_a + \hat{\sigma}^2_{d:a}/n'_d + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_r/n'_d n'_b} \quad (6)$$

$$\hat{\rho}^2(D) = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_{ab}/n'_b}{\hat{\sigma}^2_a + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_{d:a}/n'_d + \hat{\sigma}^2_r/n'_b n'_d} \quad (7)$$

Where the σ^2 terms are variance component estimates from the ANOVA and the n' terms are the numbers of levels of the subscripting facet included in the application.

The first coefficient (equation 6) is approximately equal to the expected correlation between two sets of observations on the same subjects based upon independent samples of n_d' occasions and n_b' judges. The second coefficient is approximately equal to the expected correlation between two sets observations each made on n_o' occasions by the same n_b' (fixed) set of judges. The latter coefficient is appropriate when a fixed set of observers are used in local use of the instrument. Where $n_d = n_b = 1$, these coefficients are analagous to some of the coefficient types discussed earlier. For example, the first coefficient is an interrater reliability where ratings occur on different occasions, the second is an intrarater coefficient where the same rater rates on two separate occasions. The first is an intraclass coefficient that is influenced by changes in occasions and judge when the same subjects are observed. The second reflects only changes in occasion, the difference (as can be determined from σ^2_{ab}) is the influence due to changing judges. A third and fourth coefficient of this variety could also be constructed from this design, one where occasions are fixed and one where both occasions and judges are fixed. These coefficients $\hat{\rho}^2(B)$ and $\hat{\rho}^2$ respectively are shown in equations (8) and (9). Equation (8) can be interpreted as the correlation of different judges ratings of the same subjects on the same occasion, while equation (9) is analagous to the correlation of two observations of the subjects by the same judges on the same occasions. This latter correlation is rather ab-

stract in that the data would have to be collected through videotape or some other mechanism (time sampling) that would allow observers to rate the behavior of subjects twice on the same occasion.

$$\hat{\rho}^2(B) = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_{d:a}/n'_d}{\hat{\sigma}^2_a + \hat{\sigma}^2_{d:a}/n'_d + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_r/n'_d n'_b} \quad (8)$$

$$\hat{\rho}^2 = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_{d:a}/n'_d + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_{d:ab}/n'_d n'_b}{\hat{\sigma}^2_a + \hat{\sigma}^2_{d:a}/n'_d + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_{d:ab}/n'_d n'_b + \hat{\sigma}^2_e/n'_d n'_b} \quad (9)$$

Coefficient (8) is an inter-rater coefficient where n_b' judges rate on the same occasion. Coefficient (9) is rather abstract, and has no real analogy in traditional statistics. It shows agreement of the same judge on the same occasion and is therefore an intrarater coefficient. Strictly speaking, a coefficient of this sort could be operationalized only if the observation session were videotaped.

These generalizability coefficients are by no means new to the observational literature. Mitchell (1979) speaks of three types of generalizability for observational ratings: Duplicate, Session, and Developmental. Duplicate generalizability is an

inter or intra rating of separate scorings of the same behavior and would correspond to our formulas 8 and 9 in the inter and intra cases respectively. Session generalizability focuses on scores across two occasions close in time and either coefficient (6) or (7) would be appropriate depending on the intrarater versus interrater distinction. Developmental generalizability uses the same framework except that the time frame between occasions of observation expands. The difference between session and developmental generalizability centers on whether we are charting behavior across time (developmental) or examining behavior at a single point in time (session). The distinction is analogous to the consistency vs. stability indices referred to earlier. To an even greater extent than Mitchell has suggested, generalizability theory further clarifies these distinctions as well as those associated with inter vs. intra coefficients.

Design 1. Design 1 is a variation of Design 2 where the universe of admissible observations recognizes variation due to changes in situations. Such a configuration is useful when some specified behavior is expected to change from situation to situation and this variation is of primary interest. Normally, situations would not be considered a random facet, but rather fixed at some level n_c. This facet could be considered fixed either in the universe of admissible observations, in which case a finite G study design is called for, or random in this universe and fixed in the universe of generalization. While only the latter case is considered here so that the developments remain fairly simple

from a statistical perspective, the case of fixed facets in the universe of admissible observations provides an important perspective on this topic. Four additional G coefficients, analagous to those presented in equations (6) through (9) could be constructed for this design, except that situations would be considered relevant. In this case, however, the $\hat{\rho}^2(B, C, D)$ and $\hat{\rho}^2(B, C)$ coefficients are equivalent, as are the $\hat{\rho}^2(C, D)$ and $\hat{\rho}^2(C)$ coefficients. This is because occasions are nested within the subject by situation interaction terms, and the status (fixed or random) of the occasion facet is confounded with this interaction which will never contribute to the universe score variance. The interpretation and analogy to correlation and traditional measurement statistics with all of these coefficients would be altered in the sense that the scores correlated would be considered collapsed under the same (fixed) n_b' situations. The four coefficients are indicated below:

$$\hat{\rho}^2(B,C,D) = \hat{\rho}^2(B,C) = \frac{\hat{\sigma}^2_a}{\hat{\sigma}^2_a + \hat{\sigma}^2_{ac}/n'_c + \hat{\sigma}^2_d/n'_d n'_c + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_{abc}/n'_b n'_c + \hat{\sigma}^2_r/n'_b n'_c n'_d} \tag{10}$$

$$\hat{\rho}^2(C,D) = \hat{\rho}^2(C) = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_{ab}/n'_b}{\hat{\sigma}^2_a + \hat{\sigma}^2_{ac}/n'_c + \hat{\sigma}^2_d/n'_d n'_c + \hat{\sigma}^2_{ab}/n'_b + \hat{\sigma}^2_{abc}/n'_b n'_c + \hat{\sigma}^2_r/n'_d n'_c n'_b} \tag{11}$$

Agreement. All of the above coefficients (6) through (11) are intraclass correlations which reflect the average relative rank order of two measures taken under the conditions specified by the facets of measurement. Thus, these coefficients represent reliability of rank order rather than agreement. That is, raters may consistently rank order subjects from situation to situation, or occasion to occasion, but the absolute ratings might differ, perhaps substantially. If rater agreement in the sense of score reproducibility is of interest, attention should center on error variance $\sigma^2(\Delta)$ rather than σ^2_δ which is based on $\sigma^2(\Delta)$. Error $\sigma^2(\Delta)$ reflects errors in estimating true scores and thus indicates how much variation from true score to expect on repeated observation of the same behavior. Error $\sigma^2(\Delta)$ consists of all variance components in a design, excluding those associated with true score. Computationally, the error $\sigma^2(\Delta)$ corresponding to each of the situations described in equation (6) through (11) is shown in Table 3. It makes little sense to form a coefficient based upon these errors, since the coefficient would have no interpretable analogy in classical terms. These error terms are roughly equivalent to classical theory's standard error of measurement.

Insert Table 3

About Here

Composite scores

To this point in this paper, we have not considered the possibility of composite or summed scores for decision-making. Where the sum (or average) of a set of observations is the basis upon which decisions are made, it is the "composite" score that is of interest and the observer facet is no longer explicitly considered in the universe of admissible observations. Traditionally, the question of composite scores has focused upon whether or not differential weighting (of observers) is needed or desirable. If not, the reliability of the composite rating may be computed as in the average of individual reliabilities, which is what the intraclass correlation coefficient based upon $n_{i'}$ observers indicate from a G study design. This approach is based upon the Mahalanobis distance function and is described from this perspective in Conger and Lipshitz (1973).

Both Kelly (1947) and Overall (1965) have suggested basing differential weighting schemes on the basis of the relative values of individual observer (judge) reliabilities. While Overall (1965) reports some improvement in overall reliability when this approach is used, Nagle (1952) reports little advantage, and in some cases, lowered reliabilities using this approach.

The Overall (1965) approach to determining optimal weights for observers in composite scores is to use a canonical reliability like that suggested for multivariate generalizability studies as described by Joe and Woodward (1976), and Shavelson and Webb (1981) among others. Conger and Lipshitz (1973)

describe this approach, which they attribute to Cronbach and Glaser (1953) as the determination of weights for various observers that would maximize the reliability of the composite. Other perspectives on this procedure may show more promise. If observers use different observational sets, or if the observational task is fairly difficult, this approach would prove useful to determining which observers contribute most to this maximally reliable score.

Of course, other facets could be treated in this same way. The multivariate approach could be used, for example, to detect which situations are conducive to subject discrimination for the behavior in question. Alternatively, occasions could be varied in length and number to determine optimal observation session duration and spacing.

Joe and Woodward (1976) define the multivariate coefficient as

$$\rho^2(\text{mult}) = \frac{x'\Sigma_a x}{x'\Sigma_a x + x'\Sigma_\varepsilon x(1/n_b')} \qquad (12)$$

for the single facet (B) case where the σ^2's are variance component matrices corresponding to the face of differentiation (A) and error (σ^2_δ), N_b is the number of levels of the B facet used in the D study, and x is a column vector of weights for the observer that would maximize ρ^2(mult). Coefficients which satisfy this maximum can be obtained by solving the following set of equations:

$$(\Sigma_a - \rho^2(\Sigma_a + \Sigma_\varepsilon/n_i))x = 0 \tag{13}$$

In general, this equation may be written as

$$(\Sigma_a - \rho^2(\Sigma_a + \Sigma_\delta))x = 0 \tag{14}$$

where Σ_δ is the multivariate analog to σ^2_δ in univariate generalizability theory.

The use of this framework forces us to ignore observers as an explicit facet in the design, and treats observers as multiple indicators of the same construct, some of which may contain more information regarding discrimination of subjects than others. If we have N_b observers, we may have up to N_b vectors of canonical weights, each of which describes an independent dimension of discrimination. Consequently, by examining observer characteristics (including post interview and internal rating rules) more than one dimension of discrimination may be identified, and an optimal set (number and characteristics) of raters could be determined. The presence of more than one significant dimension of discrimination may also serve as a flag that the observer rules are not being adhered to closely enough (or are ambiguous) thus suggesting revision of the procedure or instrumentation.

Focusing again on the two designs referred to earlier, the multivariate analog would treat the observer facet in each design as multiple measures of the same construct which are to be

treated as in a MANOVA, thus design 1 would be reduced to a two facet multivariate design and design 2 to a single facet multivariate design. Although generalization across fixed facets follows the same logic as that outlined earlier for univariate designs, the present discussion focuses on random effects only. Observers are considered tied to a particular design and are therefore considered fixed in the universe of admissible observations.

For design 2, a single facet design with occasions nested within subjects, the following coefficient applies

$$\rho_2^2(\text{mult}) = \frac{x'\Sigma_a x}{x'\Sigma_a x + x'\Sigma_r x} \qquad (15)$$

which may be obtained by solving

$$(\Sigma_a - \rho_2^2(\Sigma_a + \Sigma_r/n_d'))x = 0 \qquad (16)$$

where n_d' is the number of occasions nested within subjects. For design 1 the coefficient of interest is

$$\rho_1^2(\text{mult}) = \frac{x'\Sigma_a x}{x'\Sigma_a x + x'\Sigma_{ac} x/n_c' + x'\Sigma_{d,ad} x/n_d' + x'\Sigma_r x/n_c' n_d'}$$

which may be obtained by solving

$$(\Sigma_a - \rho_1^2(\Sigma_a + \Sigma_{ac}/n_c' + \Sigma_{d,ad}/n_d' + \Sigma_r/n_c' n_d'))x = 0$$

where n_d' is the number of situations in the D study design.

Illustration

More extensive data to illustrate many of the principles discussed in this paper are currently being collected. The data used in this illustration were reported in a previous study (Smith, Waller, and Waller, 1982) and should be interpreted with caution because of the small sample sizes involved. Basically, the data are classroom observation data and conform to design 2 (referred to earlier) with two observers on six occasions within nineteen teachers. Nine subscores were reported in the original study, and all were reanalyzed using concepts presented in this paper. The results for three subscores are reported here, each providing a somewhat different perspective on the analytical strategies proposed.

Table 4 shows the intrarater correlation coefficients for each rater across the three occasion pairs and the interrater correlations across all occasions within teachers. Normally, these data are the ony statistic that one might expect to be reported in a typical observational study. Table 5 shows variance component estimates for each of the five sources of variance from design 1 for each subscore. This information, in addition to generalizability coefficients would be expected to be reported if generalizability theory were used to analyze these data.

Insert Tables 4 & 5

About Here

Of the three subscores of focus here, note that the interrater correlation is high for only the first. Equation 8 where $N_d'=1$ and $n_b'=1$ (assuming occasions fixed) or Equation 6 (assuming occasions are random) would be the appropriate analogies to this correlation in generalizability theory. For the three subscores equation 8 results in values of .92, .74, and .71 respectively while equation 6 results in values .59, .54, and .23. If we were to rely only upon the interrater correlations we might conclude that subscores 2 and 3 are in need of work, but we are given no guidance as to how to improve the dependability of these subscores. For these subscores, one might wish to compute intrarater coefficients to see if raters are consistent across occasions, a requisite for a high interrater coefficient. When this is done, subscores 1 and 2 show reasonably good results across raters for the three possible rater occasion coefficients indicating that this is not likely the source of any problem. Subscore 3 shows a highly unusual pattern of correlations however, indicating little intrarater consistency for either rater. Equation 7 where $n_d'=1$ would be the generalizability theory approximation to these coefficients. They are, for the three subscores respectively, .59, .58, and .23.

The traditional and univariate generalizability theory results converge fairly well in terms of conclusions that might be drawn. The generalizability theory approach is probably more revealing, however, in that the magnitudes of the variance component estimates may be readily interpreted and highly suggestive of remedial activity for a subscore.

For example, the 'rater' component is fairly small for all three subscores indicating a high level of consistency across raters, at least in terms of mean ratings across teachers and occasions. The third subscore is a possible exception and will be discussed in more detail later in this section. Focusing on the 'O, SO' and the 'Residual' component estimates, the problem with subscore 3 become more apparent. The obvious solution here is to sample more occasions to attain the desired level of dependability, a solution not obviously offered from the correlational data presented earlier. The specific problem with this subscore is still not readily apparent, however, since the RT and TO interaction sources are confounded in the residual.

To add a final perspective on these data, the canonical solution to multivariate generalizability theory suggested by Woodward and Joe (1976) was applied to these three subscores using raters as multiple dependent variables. Table 5 shows the multivariate generalizability coefficients and scaled canonical patterns for each subscore ($n_d'=2$). The multivaritate procedure adds little to our interpretation of subscore 1. All appears well and the heavy weight on one rater and light weight on the other signifies the need for only a single rater.

Analysis of the second subscores shows two (mult) values above .70. The pattern of weights in the canonical vector indicates that each obeserver may be rating the teachers on relatively independent dimensions. This could easily explain the lower interrater coefficient, and would suggest further work on definition of the behavior to be rated.

In contrast, variable three has one multivariate coefficient greater than .70, however, the pattern vector associated with this shows roughly equal but opposite weights. While this particular result may be due to sampling error due to the small samples used, it would indicate that the observers are using opposite criteria to arrive at their ratings and that some discussion of the rating criteria needs to take place before the scale can be used dependably. While such a situation would be expected to reveal itself in the interrater correlations, it is not inconceivable that a contrast of two raters could describe the maximum differentiation of raters. Again, this would indicate that some work in trait definition or the rating procedure is warranted.

Conclusions

In this paper we have reviewed several statistical approaches to assessing the dependability of behavioral observation. It was argued that while traditional correlational statistics provide adequate information for some situations, generalizability theory removes much of the ambiguity in interpretation by explicitly defining a universe of admissible observations and a universe of generalization. Various analogies to traditional statistics were made using the statistics of generalizability theory to highlight the similarity of concept and the less ambiguous interpretation.

Multivariate approaches were suggested for use with composite scores and for situations where the covariation between levels of a facet are of interest. In the present example, observers were treated in this fashion and the results showed some promise. Additional work needs to be done in this area. For example, multivariate approaches can be used to assess the relative contribution of various situations or occasions (length, for example) of observation. In short further investigation is needed to explore this option and to consider the unique information that the multivariate analytical approach provides. As Smith and Kane (1981) have noted, the multivariate procedure may be ineffective because of the problems with sampling error. Without larger samples, work in this area is likely to be meaningless.

In this paper we have focused upon continuous type scores. It is apparent that more research is needed on agreement statistics when the nature of the task involved is classification on a nominal scale. Log-linear models or similar approaches could be considered for this purpose to bring into consideration many of the concepts of generalizability theory which have not as yet been considered with such scores.

References

Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 1966, 3-11.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 1959, 81-105.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement on partial credit. Psychological bulletin, 70, 1968, 213-220

Cone, J. D. The relevance of reliability and validity for behavioral assessment. Behavior Therapy, 8, 1977, 411-426.

Conger, J. & Lipshitz, R. Measures of reliability for profiles and test batteries. Psychometrika, 38, 1973, 411-428.

Cronbach, L. J. Test validation. In R. L. Thorndike (ed.) Educational Measurement, second edition, Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J., Gleser, G., Nanda, N., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J. and Gleser, G. C. Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press, 1965.

Crump, S. L. The estimation of variance components in analysis of vaiance. Biometrics, 2, 1946, 7-11.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 16, 1951, 407-424.

Finn, R. H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 30, 1970, 71-76.

Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educational and Psychological Measurement, 32, 1972, 255-265.

Fleiss, J. L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 1971, 378-382.

Ghiselli, E. E. Moderating effects and differential reliability and validity. Journal of Applied Psychology, 47, 1963, 81-86.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. Psychological Bulletin, 82, 1974, 278-288.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Joe, G. W. & Woodward, A. J. Some developments in multivariate generalizability theory. Psychometrika, 41, 1976, 205-218.

Kelly, T. L. Fundamentals of statistics. Cambridge, Mass.: Harvard University Press, 1947.

Lawshe, C. H. & Nagle, B. F. A note on the combination of ratings on the basis of reliability. Psychological Bulletin, 49, 1972, 17-20.

Levy, P. Generalizability studies in clinical settings. British Journal of Clinical Psychology, 13, 1974, 161-172.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 76, 1971, 365-377.

Mash, E. J. & McElwee, J. D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. Child Development, 45, 197d, 367-377.

McGaw, B., Wardrop, J. L., & Bunca, M. A. Classroom observation schemes: Where are the errors? American Educational Research Journal, 9, 1972, 13-27.

Mitchell, S. R. Interobserver observer agreement, reliability, and generalizability of data collected in observational studies. psychological Bulletin, 86, 1979, 376-390.

Overall, J. E. Reliability of composite ratings. Educational and Psychological Measurement, 25, 1965, 1011-1022.

Rowley, G. L. The reliability of observational measures. American Educational Research Journal, 13, 1976, 51-59.

Shavelson, R. & Dempsey-Atwood, N. Generalizability of measures of teaching behavior. Review of Educational Research, 46, 1976, 553-611.

Smith, P. L. & Kane, M. T. Multivariate Generalizability of Student Course Ratings. Paper presented at American Educational Research Assn., Los Angeles, April, 1981 (ERIC Doc. ED 206 649 TM 810 539).

Taplin, P. S. & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. Child Development, 44, 1973, 547-554.

Tinsley, H. E. A. & Weiss, D. J. Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22, 1975, 358-376.

Wieder, G. B. & Weiss, R. L. Generalizability theory and the coding of marital interactions. Journal of Consulting and Clinical Psychology, 48, 1980, 469-477.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley Publishing Co., 1972.

O/1175E

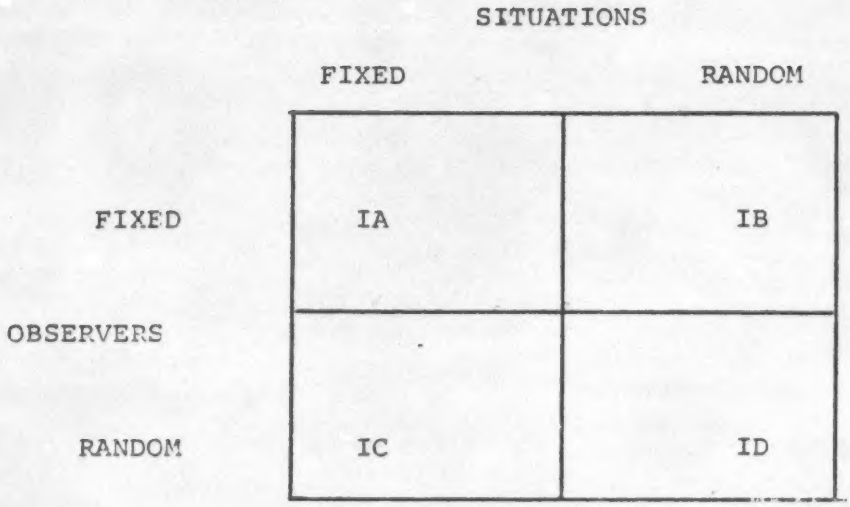

Figure 1: Four design configurations for design 1

Table 1

Expected Mean Squares and Variance Component Estimates for Design 1

Source	Expected Mean Square
Subjects(A)	$\sigma_r^2 + n\sigma_{abc}^2 + nC\sigma_{ab}^2 + nB\sigma_{ac}^2 + B\sigma_d^2 + nBC\sigma_a^2$
Situations(C)	$\sigma_r^2 + n\sigma_{abc}^2 + nA\sigma_{bc}^2 + nB\sigma_{ac}^2 + B\sigma_d^2 + nAB\sigma_c^2$
A x C	$\sigma_r^2 + n\sigma_{abc}^2 + B\sigma_d^2 + nB\sigma_{ac}^2$
Occasions w/AxC(D)	$\sigma_r^2 + B\sigma_d^2$
Observer(B)	$\sigma_r^2 + n\sigma_{abc}^2 + nA\sigma_{bc}^2 + nC\sigma_{ab}^2 + nAC\sigma_b^2$
A x B	$\sigma_r^2 + n\sigma_{abc}^2 + nC\sigma_{ab}^2$
B x C	$\sigma_r^2 + n\sigma_{abc}^2 + nA\sigma_{bc}^2$
A x B x C	$\sigma_r^2 + n\sigma_{abc}^2$
Residual	σ_r^2

$$\hat{\sigma}_a^2 = (MS_a - MS_{ac} - MS_{ab} + MS_{abc})/nBC$$

$$\hat{\sigma}_c^2 = (MS_c - MS_{ac} - MS_{bc} + MS_{abc})/nAB$$

$$\hat{\sigma}_{ac}^2 = (MS_{ac} - MS_d - MS_{abc} + MS_r)/nB$$

$$\hat{\sigma}_d^2 = (MS_d - MS_r)/B$$

$$\hat{\sigma}_b^2 = (MS_b - MS_{ab} - MS_{bc} + MS_{abc})/nAB$$

$$\hat{\sigma}_{ab}^2 = (MS_{ab} - MS_{abc})/nC$$

$$\hat{\sigma}_{bc}^2 = (MS_{bc} - MS_{abc})/nA$$

$$\hat{\sigma}_{abc}^2 = (MS_{abc} - MS_r)/n$$

$$\hat{\sigma}_r^2 = MS_r$$

Table 2

Expected Mean Squares and Variance Component Estimates for Design 2

Source	Expected Mean Squares
Subjects(A)	$\sigma_r^2 + n\sigma_{ab}^2 + B\sigma_d^2 + nB\sigma_a^2$
Occasions w/Subj.(D)	$\sigma_r^2 + B\sigma_d^2$
Observers(B)	$\sigma_r^2 + n\sigma_{ab}^2 + nA\sigma_b^2$
A x B	$\sigma_r^2 + n\sigma_{ab}^2$
Residual	σ_r^2

$$\hat{\sigma}_a^2 = (MS_a - MS_d - MS_{ab} + MS_r)/nB$$

$$\hat{\sigma}_d^2 = (MS_d - MS_r)/B$$

$$\hat{\sigma}_b^2 = (MS_b - MS_{ab})/nA$$

$$\hat{\sigma}_{ab}^2 = (MS_{ab} - MS_r)/n$$

$$\hat{\sigma}_r^2 = MS_r$$

Table 3

Error $\hat{\sigma}^2(\Delta)$ Associated with the Designs

Characterized by Equations 6 through 11

Equation	$\sigma^2(\Delta)$
6	$\sigma_d^2/n_d' + \sigma_b^2/n_b' + \sigma_{ab}^2/n_b' + \sigma_r^2/n_d'n_b'$
7	$\sigma_d^2/n_d' + \sigma_r^2/n_d'n_b'$
8	$\sigma_b^2/n_b' + \sigma_{ab}^2/n_b' + \sigma_r^2/n_d'n_b'$
9*	$\sigma_r^2/n_d'n_b'$
10	$\sigma_c^2/n_c' + \sigma_{ac}^2/n_c' + \sigma_d^2/n_d'n_c' + \sigma_b^2/n_b' + \sigma_{ab}^2/n_b' + \sigma_{bc}^2/n_b'n_c' + \sigma_{abc}^2/n_b'n_c' + \sigma_r^2/n_d'n_b'n_c'$
11	$\sigma_c^2/n_c' + \sigma_{ac}^2/n_c' + \sigma_d^2/n_d'n_c' + \sigma_{bc}^2/n_b'n_c' + \sigma_{abc}^2/n_b'n_c' + \sigma_r^2/n_d'n_b'n_c'$

Table 4

Interobserver and Intraobserver Correlations for

Subtests 1,2, and 3

	Intraobserver (rater 1/2)			Interobs.
Subscore	occ1-occ2	occ1-occ3	occ2-occ3	
1	.37/.36	.71/.81	.77/.68	.94
2	.17/.81	.25/.85	.90/.70	.76
3	-.08/-.31	.66/-.18	.43/.41	.73

Table 5

Variance Component Estimates for

Subtests 1,2, and 3

Var. Comp.	Subtest			
	1	2		3
σ_a^2	52.86	24.98		2.63
σ_b^2	0.00	.79		2.33
σ_d^2	29.78	9.43		5.69
σ_{ab}^2	0.00	1.98		0.00
σ_r^2	6.94	10.09		3.31
ρ^2(mult)	.77	.88	.76	.72
Scaled Vector	.074	.13	-.05	.56
	.049	-.03	.16	-.80